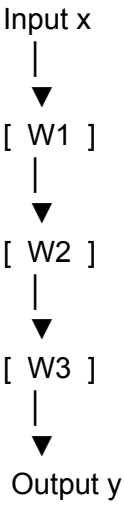
Here's a clean diagram contrasting **purely linear stacking** vs **ReLU-gated (nonlinear) stacking**.

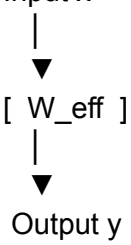
🔷 1. Deep Linear Network (Collapses)



Collapse:

$W_{eff} = W_3 \cdot W_2 \cdot W_1$

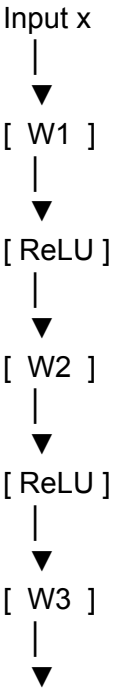
So the entire network becomes:



Key Point:

➡️ Only **one fixed linear map**
➡️ Capacity = capacity of a single matrix
➡️ Depth adds **no new expressive power**

🔷 2. ReLU-Gated Network (Does NOT Collapse)



🔁 Rewrite using gating matrices

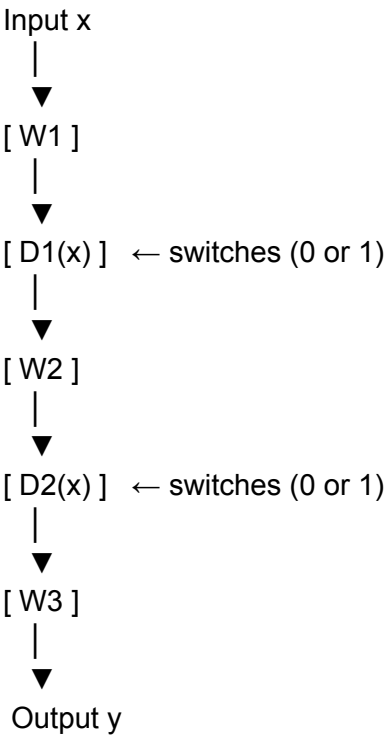
Each ReLU becomes a diagonal switch matrix:

$$\text{ReLU}(z) = D(z) \cdot z$$

So the network becomes:

$$y = W3 \cdot D2(x) \cdot W2 \cdot D1(x) \cdot W1 \cdot x$$

🔷 3. Effective Structure



🔷 4. Key Difference (Side-by-Side)

Linear case:

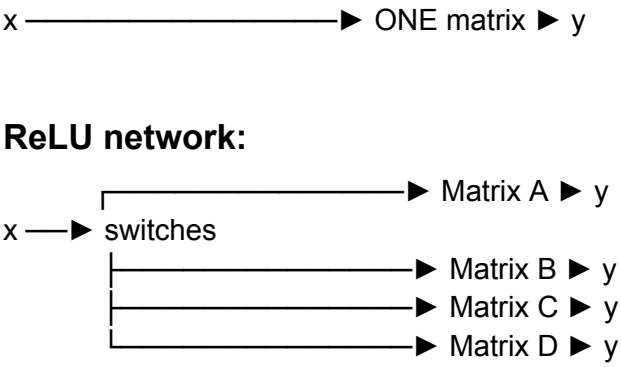
$W3 \cdot W2 \cdot W1 \rightarrow$ single fixed matrix

ReLU case:

$W3 \cdot D2(x) \cdot W2 \cdot D1(x) \cdot W1 \rightarrow$ depends on x values at each layer.

🔷 5. Big Insight (Visual Intuition)

Linear network:



➡️ The input **selects a path**
➡️ Each path = a different effective matrix

🔷 6. Interpretation

- **Linear stack**
  → one associative memory
- **ReLU network**
  → exponentially many **conditional associative memories**

🔷 7. Compact Summary Diagram

LINEAR:
x → W1 → W2 → W3 → y
↓ collapse
x → W_eff → y

NONLINEAR (ReLU):
x → W1 → D1(x) → W2 → D2(x) → W3 → y
↓
input-dependent routing
↓
many possible W_eff(x)